Facial trustworthiness and criminal sentencing: A comment on Wilson and Rule (2015)

Robin S. S. Kramer and Ellen M. Gardner

School of Psychology, University of Lincoln

Corresponding Author:

Robin Kramer, School of Psychology, University of Lincoln, Lincoln, LN6 7TS, UK.

Email: remarknibor@gmail.com

Telephone: +44 (0)1522 882 000

Abstract

Our first impressions of others, whether accurate or unfounded, have real-world consequences in terms of how we judge and treat those people. Previous research has suggested that criminal sentencing is influenced by the perceived facial trustworthiness of defendants in murder trials. In real cases, those who appeared less trustworthy were more likely to receive death rather than life sentences. Here, we carried out several attempts to replicate this finding, utilising the original set of stimuli (Study 1), multiple images of each identity (Study 2), and a larger sample of identities (Study 3). In all cases, we found little support for the association between facial trustworthiness and sentencing. Further, there was clear evidence that the specific image chosen to depict each identity had a significant influence on subsequent judgements. Taken together, our findings suggest that perceptions of facial trustworthiness have no real-world influence on sentencing outcomes in serious criminal cases.

Keywords

Face perception, trustworthiness, criminal sentencing, first impressions, death sentence

Short title

Facial trustworthiness and criminal sentencing

Introduction

Our first impressions of others are often based on minimal information and yet their formation may be fast and automatic (Hassin & Trope, 2000; Ritchie, Palermo, & Rhodes, 2017; Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006). Research has shown that these initial perceptions often generalise, resulting in assumptions regarding additional judgements. For instance, attractive individuals are also assumed to possess more socially desirable personality traits (Dion, Berscheid, & Walster, 1972). While recent evidence suggests that judgements based solely on facial appearance provide some level of predictive validity with respect to personality and other behavioural measures (e.g., Kramer & Ward, 2010, 2011; Little & Perrett, 2007; Penton-Voak, Pound, Little, & Perrett, 2006), it remains the case that our impressions may often also be unfounded (Olivola & Todorov, 2010).

Although the origins of trait inferences are currently unknown, some researchers suggest that they are caused by an overgeneralisation of emotion recognition systems – we misattribute traits based on a neutral face's subtle resemblance to emotional expressions (Said, Sebe, & Todorov, 2009). Other evidence supports an account whereby facial cues signalling approach/avoidance and physical strength/weakness are overgeneralised, resulting in our perceptions of valence and dominance respectively (Oosterhof & Todorov, 2008).

Although often characterised as innate, the recent 'Trait Inference Mapping' framework (Over & Cook, 2018) argues that these inferences are due to learned mappings between 'face space' and 'trait space'. The formation of internal spaces in which we separately represent faces and traits takes place through experience, along with a learned mapping between locations in the two spaces for each encountered identity. As a result, the location of an unfamiliar person in face space is used to infer their traits using our prior knowledge regarding mappings between the two spaces. Whether this or another account proves better in explaining their underlying mechanisms, it is clear that trait inferences are both common and influential.

Despite the possibility that facial appearance may not predict behaviour, studies using hypothetical scenarios have demonstrated that a defendant's baby-facedness (Berry & Zebrowitz-McArthur, 1988), perceived trustworthiness (Korva, Porter, O'Connor, Shaw, & ten Brinke, 2013), perceived attractiveness (Desantts & Kayson, 1997; Wuensch, Castellow, & Moore, 1991), facial expression (Abel & Watters, 2005), and facial tattoos (Funk & Todorov, 2013) influenced their perceived guilt. Importantly, however, such biases are not guaranteed to be present in actual courtroom judgements.

Addressing this question of ecological validity, facial first impressions have also been found to have measurable influences on real-world outcomes. For example, perceptions of facial trustworthiness and competence are associated with election outcomes (Chen, Jing, & Lee, 2014; Todorov, Mandisodza, Goren, & Hall, 2005),

baby-facedness shows a relationship with adjudications in small claims court (Zebrowitz & McDonald, 1991), and Afrocentric appearance has been linked with criminal sentencing (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006). General (rather than specifically facial) attractiveness has also been shown to influence hiring probability and wages (Pfeifer, 2012), as well as criminal sentencing (Downs & Lyons, 1991; Stewart, 1980, 1985).

In recent years, an already well-cited article by Wilson and Rule (2015) argued that the perceived trustworthiness of convicted criminals' faces was able to predict their sentencing outcomes in real-world cases. Specifically, those rated as less trustworthy were more likely to have received a death (rather than life) sentence. In their first study, the stimulus sample comprised inmates incarcerated by the Florida Department of Corrections, either serving life sentences or awaiting execution. Importantly, and as the authors note, evidence has shown that facial appearance may actually reflect a person's aggression and trustworthiness (e.g., Carré, McCormick, & Mondloch, 2009; Stirrat & Perrett, 2010), and so criminals who appeared less trustworthy might also be have been more violent, and therefore deserving of harsher sentences. In other words, facial appearance may not, of itself, have influenced sentencing decisions.

Wilson and Rule (2015) aimed to address this issue with their second study by utilising a stimulus set comprising only innocent men who were later exonerated. As such, any association between perceived trustworthiness and sentencing outcomes could

not have been mediated by the individual's actual behaviours (e.g., greater violence). The results of the study demonstrated that trustworthiness judgements also predicted sentencing (i.e., whether the defendant received a life or death sentence) in these innocent people. The researchers therefore concluded that facial appearance alone, and not associated behaviours, predicted harsher sentences, presumably through its influence on judges and members of the jury.

Importantly, the facial judgements collected in this second study were based upon unconstrained photographs, taken from the biographical profiles featured on the 'Innocence Project' website. As such, these images varied in expression, pose, lighting, resolution, clothing worn, distance to camera, and so on. It is also worth noting that the photographs were not taken during the trial, instead depicting the men at an undetermined point in their lives (and appearing to include images taken before serving time in prison for some men and after serving time for others). Studies have shown that these image factors influence trait impressions, and that the same person is perceived as higher or lower in trustworthiness, for example, depending upon the particular image chosen (Jenkins, White, Van Montfort, & Burton, 2011; Todorov & Porter, 2014). Indeed, a simple smile will have measurable effects on trustworthiness perceptions (e.g., Schmidt, Levenstein, & Ambadar, 2012). Therefore, the finding that judgements based upon a single, unconstrained image of each person were able to predict sentencing

decisions is surprising, considering also that these images did not depict the individuals as they appeared in court (and hence as judges and juries viewed them).

To explore this further, we first attempted to replicate the original finding that trustworthiness ratings of the faces of innocent men predicted the sentences that they received (Study 2; Wilson & Rule, 2015). Next, given that trait impressions are known to be image-dependent (Jenkins et al., 2011; Todorov & Porter, 2014), we considered how image choice might influence the trustworthiness–sentencing relationship. Finally, we attempted to replicate Wilson and Rule's (2015) original design using a newly collected stimulus set. To anticipate the results, we found little evidence to support the association between perceived facial trustworthiness and sentencing outcomes.

Study 1 – Replication of Wilson and Rule's (2015) study

In this first study, we replicated the second experiment of Wilson and Rule (2015), using the same stimuli as in their study while recruiting a larger sample of raters. By focussing on the sentencing of innocent men, we could be more confident that perceptions of trustworthiness alone, and not the actual behaviours of the men, were influencing sentencing decisions.

Method

Participants

A sample of 103 American volunteers (age *M* = 35.67 years, *SD* = 12.01 years; 40% women; 86% self-reported as White) gave informed, onscreen consent before participating in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited through Amazon Mechanical Turk (MTurk), a crowdsourcing website that allows 'workers' to complete online tasks. Previous research has established MTurk as a reliable source of data (e.g., Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). This experiment and those below were approved by the University of Lincoln's School of Psychology ethics committee (PSY1718564) and were carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. There was no overlap between this sample and those who participated in Studies 2 and 3.

Stimuli

The 37 greyscale images of men's faces featured in Wilson and Rule's (Study 2; 2015) experiment were used here (having been deposited online by the authors). Following Wilson and Rule, these images were presented at approximately 118 x 118 pixels.

Procedure

The experiment was completed online using the Qualtrics survey platform (www.qualtrics.com). After consent was obtained, participants provided demographic information (age, sex, and ethnicity).

Following Wilson and Rule (2015), each participant was asked to rate all 37 images, presented in a random order, for trustworthiness. The instruction presented onscreen throughout the experiment read, "How trustworthy would you rate the person in this image?" (1 = not at all trustworthy, 8 = very trustworthy). Cronbach's α for interrater reliability was .87.

Results and Discussion

In Wilson and Rule's (2015) original study, using the 37 images presented here, ratings were averaged across their 39 participants (also recruited via MTurk) for each image. Subsequently, a *by-stimulus* analysis was carried out, meaning that their results could be generalised to other samples of stimuli but not other participants (Judd, Westfall, & Kenny, 2012). Their logistic regression found that trustworthiness significantly

predicted sentencing outcomes (coding death as 0 and life as 1), $b = 1.55$, $SE = 0.68$, $p = .022$; odds ratio (OR) = 4.71, 95% CI [1.25, 17.76].

Our logistic regression, following the same process, found that trustworthiness was not a significant predictor of sentencing outcomes, $b = 0.85$, $SE = 0.96$, $p = .375$, OR = 2.34, 95% CI [0.36, 15.29]. In addition, we carried out a *by-participant* analysis, calculating a regression coefficient for each participant separately (i.e., predicting sentencing outcomes using that participant's trustworthiness ratings only) and then comparing these coefficient values to zero. This type of analysis, common in research on social judgements (e.g., Huang et al., 2018), produces results that generalise to other samples of participants but not to other samples of stimuli (Judd et al., 2012). Although the individual coefficients were significantly different from zero, $t(102) = 2.04$, $p = .044$, Cohen's $d = 0.20$, the mean coefficient across participants was small and suggested little predictive strength, $b = 0.09$, [0.00, 0.17].

Taken together, these results demonstrate that facial trustworthiness has little (although statistically different from zero) ability to predict sentencing outcomes for this set of stimuli, a pattern that we would likely find using other samples of participants. However, at the stimulus level, we found no evidence that trustworthiness predicted sentencing outcomes for these images or would do so for other images (i.e., other innocent men who received life or death sentences).

Study 2 – Ratings of multiple photographs of each identity

In this second study, we investigated the importance of the particular images used to depict the men since previous research has demonstrated that trait impressions are known to be image-dependent (Jenkins et al., 2011; Todorov & Porter, 2014). By collecting multiple images for each identity, we explored how image choice affected subsequent judgements and the predictive value of trustworthiness perceptions.

Method

Participants

A sample of 138 British volunteers (age $M = 24.68$ years, $SD = 9.61$ years; 75% women; 94% self-reported as White) gave informed, onscreen consent before participating in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited through the university's SONA system (receiving course credit for participation), MTurk, and by word of mouth (sharing the experiment's weblink on social media). There was no overlap between this sample and those who participated in Studies 1 and 3.

The data from 20 additional participants were excluded before analyses as they failed to complete the task, most likely due to the length of the study (rating 255 images) in comparison with Studies 1 (37 images) and 3 (44 images).

Stimuli

For the 37 men used in Wilson and Rule's (2015) study, we collected all available photographs (avoiding duplications) using Google Images searches with the person's name. This resulted in a set of 255 images, varying in the number of images per person (M = 6.89 images, SD = 3.91 images; range = 2-16 images). All images were cropped to show just the faces and minimal background (118 x 118 pixels, following Wilson & Rule, 2015), and featured images in colour where available. Importantly, this set included the original (greyscale) images used by Wilson and Rule.

Procedure

The experiment was completed online using the Qualtrics survey platform. After consent was obtained, participants provided demographic information (age, sex, and ethnicity).

Following Wilson and Rule (2015), each participant was asked to rate all 255 images, presented in a random order, for one of four traits. Participants were randomly assigned to traits initially, although towards the end of data collection, we chose to recruit a larger sample of trustworthiness ratings (in line with Wilson & Rule, 2015) in order to provide additional power for our analyses focussing on this trait.

The instruction presented onscreen throughout the experiment read, "How X would you rate the person in this image?" where X referred to "trustworthy" (1 = not at all trustworthy, 8 = very trustworthy), "attractive" (1 = not at all attractive, 8 = very attractive), and "mature-faced" (1 = baby face, 8 = mature face). For Afrocentricity, we provided a more detailed description, "How much does the person in this image show Afrocentric features (features that are more typical of an African American e.g. skin colour, hair, eyes, nose, eyes, cheeks, lips etc.)?", along with the accompanying scale (1 = not at all Afrocentric, 8 = very Afrocentric). A summary of the sample sizes and interrater reliabilities for the traits can be seen in Table 1.

Table 1. A summary of the sample sizes and interrater reliabilities for the four traits.

Trait	*N*	Cronbach's α
Trustworthiness	56	.90
Attractiveness	27	.87
Facial maturity	25	.94

Afrocentricity	30	.99

Results and Discussion

Following Wilson and Rule (2015), we calculated the mean rating for each image for each trait separately, and as such, carried out by-stimulus analyses.

First, we considered the mean ratings for the original 37 images only (i.e., those used in the original study). Mirroring Wilson and Rule's approach, our first step employed a logistic regression model in order to determine whether trustworthiness ratings predicted sentence outcomes (0 = death, 1 = life). We found that trustworthiness was not a significant predictor (p = .116), and the model did not account for more variance than the intercept-only model, $\Delta\chi^2(1) = 2.73$, p = .099.

In a second step, we entered all other covariates (Afrocentricity, attractiveness, facial maturity, the presence of glasses, and time served) included in Wilson and Rule's analysis. In line with their approach, given that Afrocentricity ratings were distributed bimodally according to race, mean Afrocentricity ratings for the images were normalised within White and Black identities separately (i.e., for each race, the mean image ratings were transformed, giving M = 0, SD = 1). The addition of these covariates did not improve the model, $\Delta\chi^2(5) = 6.72$, p = .242. The results of both models can be seen in Table 2.

Table 2. Results of the logistic regression analyses predicting sentence outcome (death = 0, life = 1) for the original 37 images.

Predictor		*b*	OR
Model 1			
	Trustworthiness	1.20 (0.76)	3.32 [0.74, 14.80]
	Intercept	-5.83 (3.83)	0.00
Model 2			
	Trustworthiness	0.96 (0.93)	2.60 [0.42, 16.22]
	Afrocentricity	0.37 (0.44)	1.45 [0.62, 3.42]
	Attractiveness	0.15 (0.90)	1.16 [0.20, 6.70]
	Facial maturity	0.26 (0.43)	1.30 [0.56, 3.01]
	Presence of glasses	-1.21 (1.03)	0.30 [0.04, 2.23]
	Time served	0.15* (0.07)	1.16 [1.01, 1.34]
	Intercept	-8.20 (5.29)	0.00

Note: OR = odds ratio. Standard errors are given in parentheses; 95% confidence intervals are given in brackets. *$p < .05$.

For the by-participant analysis of trustworthiness ratings, one participant's data were excluded because they gave the same response to all except one of the images,

preventing a regression from being carried out. For the remaining participants, our analysis showed that individual coefficients did not differ from zero, $t(54) = 1.07$, $p = .289$, Cohen's $d = 0.14$, with mean coefficient $b = 0.05$, [-0.04, 0.15]. As noted earlier, this suggests that other samples of participants would also likely show no predictive effect of trustworthiness for these stimuli.

Next, we considered the dependence of the by-stimulus regression coefficient on the specific images chosen. For each of 10,000 iterations, we selected a random image for each of the 37 identities. As such, there was always one image representing each identity in each iteration. Using the mean trustworthiness ratings for this set of images, we calculated the regression coefficient. We found that 62% of iterations produced a value less than the one resulting from the original stimuli ($b = 1.20$, above), highlighting that 1) any relationship between ratings and sentencing outcomes is dependent upon the image chosen to depict each identity; and 2) the original stimuli used in Wilson and Rule's (2015) study may have overestimated the expected predictive strength of facial images for these identities.

Study 3 – Collection of a new sample from the Innocence Project website

In our final study, we decided to replicate Wilson and Rule's (2015) design using a newly collected stimulus set. The original stimuli represented all suitable identities

listed on the Innocence Project website as of October 2014. We were therefore interested to test whether trustworthiness perceptions were associated with sentencing outcomes in a larger, more recent sample of identities that were chosen to fulfil the original criteria specified by Wilson and Rule.

Method

Participants

A sample of 104 American volunteers (age $M = 33.89$ years, $SD = 9.46$ years; 34% women; 75% self-reported as White) gave informed, onscreen consent before participating in the experiment and were provided with an onscreen debriefing upon completion. Participants were recruited through MTurk. There was no overlap between this sample and those who participated in Studies 1 and 2.

Stimuli

We collected information for every man listed on the Innocence Project website (www.innocenceproject.org) whose biographical profile contained a photograph (244 men as of March 2019). Two of the original profiles featured in Wilson and Rule's

study no longer provided photographs and so were not included here. For the 244 men identified, we recorded the man's sentence received, how long he served, the crime(s) he was convicted of, the year of conviction, and in which state the conviction occurred.

We specifically followed Wilson and Rule's two exclusion criteria regarding identity selection, choosing to be conservative where additional decisions were necessary. As a result, from this list, we first eliminated any man who lived in a state in which the death sentence was not administered at the time of conviction (including three identities featured in the original study).

Second, we selected only men whose crimes would have made them eligible for the death penalty in their states. This criterion eliminated one identity included in the original study, who was not charged with murder. We chose to include only those who were convicted of murder (including first degree murder, aggravated murder, capital murder) as these crimes are eligible to receive the death penalty. In contrast, we excluded those convicted of felony murder or second degree murder as these crimes are not typically eligible to receive a death sentence.

Finally, we also chose to select only those men that received a death or life sentence. It was unclear how this was defined in the original study, given that Wilson and Rule included men who had received both life sentences (either life or life without parole) and those sentenced to a specific number of years (e.g., 99 years). Rather than selecting an arbitrary cut off for the number of years that might be equivalent to a life

sentence, we opted to include only those men that were sentenced to life, life without parole, or life + X (where X represents some additional number of years on top of their life sentence).

These criteria resulted in a final set of 44 targets: 17 sentenced to death and 27 sentenced to life. Of these targets, 23 were Black and 21 were White or Hispanic. Of our 44 targets, 27 appeared in the original set of identities although only 11 were depicted using the same image. (No reason was given for why profile images had been replaced on the website.)

The images were converted to greyscale and cropped as above (118 x 118 pixels) to show only the faces.

Procedure

The procedure was identical to that used in Study1, with each participant rating 44 images. Cronbach's α for interrater reliability was .88.

Results and Discussion

As above, we calculated the mean trustworthiness rating for each image. Following Wilson and Rule's (2015) consideration of race for the original set of identities, we first

investigated whether race was associated with trustworthiness for this new sample of men. In a 2 (Race: Black, White) × 2 (Sentence: death, life) between-subjects ANOVA, there was no main effect of sentence, $F(1, 40) = 2.05$, $p = .160$, $\eta_p^2 = .05$, or race, $F(1, 40) = 0.64$, $p = .429$, $\eta_p^2 = .02$, on trustworthiness ratings and no race x sentence interaction, $F(1, 40) = 0.05$, $p = .827$, $\eta_p^2 = .00$. Therefore, race was not considered in subsequent analyses.

In a by-stimulus regression, we found that trustworthiness was not a significant predictor of sentencing outcomes for these men, $b = 1.23$, $SE = 0.86$, $p = .154$, OR = 3.41, 95% CI [0.63, 18.40].

For the by-participant analysis, three participants' data were excluded because they gave the same response to all of the images, preventing a regression from being carried out. For the remaining participants, our analysis showed that individual coefficients differed from zero, $t(100) = 4.13$, $p < .001$, Cohen's $d = 0.41$, with a small mean coefficient $b = 0.11$, [0.06, 0.16].

Mirroring the results of Study 1, facial trustworthiness has little (although statistically different from zero) ability to predict sentencing outcomes for this set of stimuli, a pattern that should generalise to other samples of participants. However, at the stimulus level, there was no evidence that trustworthiness predicted sentencing outcomes for these images/identities, and we would expect this to be the case for other stimuli also.

General Discussion

In a recent study, Wilson and Rule (2015) provided evidence that the perceived facial trustworthiness of innocent men (where behavioural correlates could not play a role) predicted their sentencing outcomes in real murder trials. Here, across three studies, our findings fail to support this conclusion.

In a direct replication of the previous study, using the same set of stimuli, our ratings of trustworthiness did not predict sentencing outcomes when analysed at the level of the image. While participant-level analyses produced regression coefficients that were greater than zero, these remained small and provided little predictive strength.

In our second study, we demonstrated that judgements of facial trustworthiness were image-dependent, suggesting that ratings collected using a single, unconstrained photograph of each identity must inherently provide idiosyncratic results. To show (as Wilson and Rule, 2015, did) that a specific set of images predicts sentencing outcomes is therefore hard to interpret since these particular images cannot be taken as the equivalent of the identities being judged. By demonstrating how image choice affected prediction outcomes, we provided a clear argument against generalising results for a single image set to the identities under consideration. In addition, we again failed to

replicate the association between trustworthiness and sentencing for the original image set.

In Study 3, we aimed to test the relationship between perceptions and sentencing for a larger set of identities. To this end, we collected new stimuli (although overlap with the original set of men was to be expected) and our ratings of trustworthiness again failed to predict sentencing outcomes. As with Study 1, participant-level analyses produced regression coefficients that were small (although greater than zero) and provided little predictive strength.

Perhaps the clearest criticism of Wilson and Rule's (2015) study must be levelled at the nature of the stimuli used. By collecting trait impressions based upon unconstrained images that featured the men at times other than during their trials, it is hard to see how such impressions could provide a proxy for those formed by the judges and jurors. In contrast, if Wilson and Rule were to argue that the use of any image results in impressions representative of those formed by members of the court, then this simply is not the case (Jenkins et al., 2011; Todorov & Porter, 2014). That the previous work found a relationship between perceived facial trustworthiness and sentencing despite this issue is surprising, and likely explains why such a result was not replicated in the current set of studies.

Taken together, our results suggest no reason to consider facial judgements as predictive of life and death sentencing outcomes. At best, the original findings were

overstated, with the relationship between trustworthiness and sentencing either absent or at least insignificant in terms of real-world effects. Indeed, the experimental design itself (here and in the original study) is insufficient in demonstrating a causal relationship even if an association was present. Any compelling evidence that facial appearance prejudices perceivers to the extent that it influences life and death decisions would require a far more in-depth examination of the process and those involved.

Going beyond the experimental evidence, we propose that numerous crucial factors would have overshadowed any potential influence of facial appearance in these cases. For example, two of the men in the stimulus set pled guilty and testified against a supposed accomplice in order to receive life rather than death sentences. Indeed, research has confirmed that the threat of the death penalty increases the likelihood of a plea agreement (Thaxton, 2013). In such cases, it seems reasonable to assume that outcomes were not the result of a more trustworthy face. Researchers have also identified other factors that play a significant role in death penalty decisions. For example, defendants are more likely to be sentenced to death by judges seeking (re)election (Brooks & Raphael, 2001; Canes-Wrone, Clark, & Kelly, 2014), although in the majority of cases, it is the jury that is responsible for deciding whether a death sentence is given.

In contrast, comparatively more minor legal decisions, such as those made in small claims court, may show an influence of trait impressions since judgements depend

largely on the credibility of litigants, who typically have little evidentiary support (Zebrowitz & McDonald, 1991). In these types of cases, there may be more opportunity for judges to be swayed by potentially unconscious biases. For more serious crimes, sentencing constraints (e.g., evidence of aggravating or mitigating factors) may limit the influence of these biases. However, further research is needed in order to investigate the veracity of this idea.

In conclusion, the current studies fail to replicate the association between facial trustworthiness and criminal sentencing that has previously been demonstrated (Wilson & Rule, 2015). As such, we recommend that researchers show caution when investigating this topic in future by questioning this relationship and pursuing more exacting tests of the hypothesis using more appropriate experimental designs. For instance, constrained (passport-style) images depicting defendants during their trials would allow researchers to better investigate trustworthiness impressions that are formed by judges and juries, and how these might influence sentencing outcomes.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interest

The authors declare that there is no conflict of interest.

References

Abel, M. H., & Watters, H. (2005). Attributions of guilt and punishment as functions of physical attractiveness and smiling. *The Journal of Social Psychology, 145*(6), 687-702.

Berry, D. S., & Zebrowitz-McArthur, L. (1988). What's in a face? Facial maturity and the attribution of legal responsibility. *Personality and Social Psychology Bulletin, 14*(1), 23-33.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*(10), 674-679.

Brooks, R. R. W., & Raphael, S. (2001). Life terms or death sentences: The uneasy relationship between judicial elections and capital punishment. *Journal of Criminal Law & Criminology, 92*(3-4), 609-640.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5.

Canes-Wrone, B., Clark, T. S., & Kelly, J. P. (2014). Judicial selection and death penalty decisions. *American Political Science Review, 108*(1), 23-39.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science, 20*(10), 1194-1198.

Chen, F. F., Jing, Y., & Lee, J. M. (2014). The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology, 51*, 27-33.

Desantts, A., & Kayson, W. A. (1997). Defendants' characteristics of attractiveness, race, and sex and sentencing decisions. *Psychological Reports, 81*(2), 679-683.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology, 24*(3), 285-290.

Downs, A. C., & Lyons, P. M. (1991). Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin, 17*(5), 541-547.

Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., & Johnson, S. L. (2006). Looking deathworthy: Perceived stereotypicality of Black defendants predicts capital-sentencing outcomes. *Psychological Science, 17*(5), 383-386.

Funk, F., & Todorov, A. (2013). Criminal stereotypes in the courtroom: Facial tattoos affect guilt and punishment differently. *Psychology, Public Policy, and Law, 19*(4), 466-478.

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*(5), 837-852.

Huang, J., He, X., Ma, X., Ren, Y., Zhao, T., Zeng, X., ... & Chen, Y. (2018). Sequential biases on subjective judgments: Evidence from face attractiveness and ringtone agreeableness judgment. *PLoS ONE, 13*(6), e0198723.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54-69.

Korva, N., Porter, S., O'Connor, B. P., Shaw, J., & ten Brinke, L. (2013). Dangerous decisions: Influence of juror attitudes and defendant appearance on legal decision-making. *Psychiatry, Psychology and Law, 20*(3), 384-398.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology, 63*(11), 2273-2287.

Kramer, R. S. S., & Ward, R. (2011). Different signals of personality and health from the two sides of the face. *Perception, 40*(5), 549-562.

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology, 98*, 111–126.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315-324.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092.

Over, H., & Cook, R. (2018). Where do spontaneous first impressions of faces come from? *Cognition, 170*, 190-200.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411-419.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition, 24*, 490–524.

Pfeifer, C. (2012). Physical attractiveness, employment and earnings. *Applied Economics Letters, 19*(6), 505-510.

Ritchie, K. L., Palermo, R., & Rhodes, G. (2017). Forming impressions of facial attractiveness is mandatory. *Scientific Reports, 7*(1), 469.

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*(2), 260-264.

Schmidt, K., Levenstein, R., & Ambadar, Z. (2012). Intensity of smiling and attractiveness as facial signals of trustworthiness in women. *Perceptual and Motor Skills, 114*(3), 964-978.

Stewart, J. E., II. (1980). Defendant's attractiveness as a factor in the outcome of criminal trials: An observational study. *Journal of Applied Social Psychology, 10*(4), 348-361.

Stewart, J. E., II. (1985). Appearance and punishment: The attraction-leniency effect in the courtroom. *The Journal of Social Psychology, 125*(3), 373-378.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust: Male facial width and trustworthiness. *Psychological Science, 21*(3), 349-354.

Thaxton, S. (2013). Leveraging death. *Journal of Criminal Law & Criminology, 103*(2), 475-552.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308*(5728), 1623-1626.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27*(6), 813-833.

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science, 25*(7), 1404-1417.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7), 592-598.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*(8), 1325-1331.

Wuensch, K. L., Castellow, W. A., & Moore, C. H. (1991). Effects of defendant attractiveness and type of crime on juridic judgment. *Journal of Social Behavior and Personality, 6*(4), 713-724.

Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior, 15*(6), 603-623.